



COURSERA FINAL PROJECT

SPECIALIZED MODELS: TIME SERIES AND SURVIVAL ANALYSIS

# Time Series Forecasting to predict grocery sales at Favorita stores

Project presented by

**Raziel Amador Rios**

to obtain the Coursera Certificate

**Main object:** produce a reliable forecast based on sales time-series data from a retail store (Favorita stores, an Ecuadorian-based company).

**Barcelona, February 2023**



---

# Table of Contents

<b>Table of Contents</b>	<b>iii</b>
<b>Time Series Forecasting</b>	<b>v</b>
I Main objective . . . . .	vi
II Data description . . . . .	vii
III Data exploration and data cleaning . . . . .	viii
III.1 Time series Decomposition . . . . .	ix
III.2 Stationarity and seasonality analyses . . . . .	x
<b>Appendix</b>	<b>xi</b>
IV Supplementary information . . . . .	xii



---

# Time Series Forecasting

## I

## Main objective

The **main object** of the project is to:

- Generate a reliable **time series forecast** based on sales from Favorita stores.

Favorita stores is a grocery retail store based in Ecuador. Producing an accurate forecast could lead to decreased food waste related to overstocking and improve customer satisfaction. All the code can be found in the following [GitHub repository](#), further information can be found in Supplementary information.

## II

## Data description

The dataset comes from **Kaggle**, named: [Store Sales - Time Series Forecasting](#). The Kaggle dataset contains sales data from *Corporación Favorita*, a large Ecuadorian-based grocery retailer. Kaggle provides you with 7 different files (see Table 1 for more details).

Number	File Name	Description
1	holiday_events.csv	Relevant holidays in Ecuador
2	oil.csv	Oil prices from 2013 to 2017
3	sample_submission.csv	submission example
4	stores.csv	Stores metadata
5	test.csv	Stores and family products
6	train.csv	Sales by store and product-family from 2013-01 to 2017-08
7	transactions.csv	Number of transactions by store

**Table 1: Kaggle files description.** Files are alphabetically sorted.

The training data represents 99% of the data, including dates from 2013-01-01 to 2017-08-16 (55.5 months), 54 stores placed in different cities within Ecuador, and 33 family-products (see Figure 1). The testing data includes dates from 2017-08-16 to 2017-08-31 (15 days).

Stores Summary					
<b>54</b> Stores	<b>5</b> Store types	<b>17</b> Store clusters	<b>33</b> Product families	<b>16</b> States	<b>55.5</b> Months

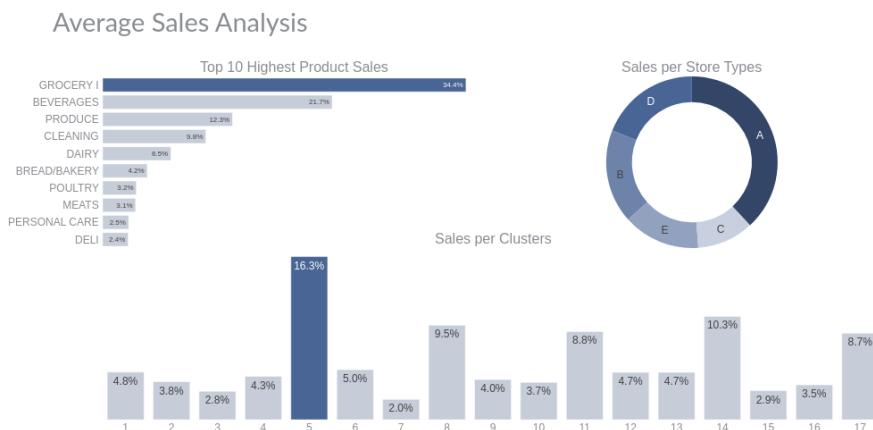
**Figure 1: Summary of the training dataset.** The cluster information denotes similarity between stores.

## III

---

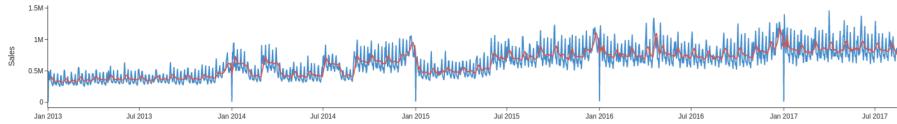
## Data exploration and data cleaning

Performing an exploratory data analysis (eda) of the sales from the training dataset, we can observe that grocery I, beverages, and produce are the top 3 most consumed products (see Figure 2). Additionally, store type A, and cluster 5 are the most frequent among their classification (Figure 2).



**Figure 2: EDA of sales.** The plot describes the sales by product, store type, and per cluster (left, right, and below, respectively). Darker blue represents higher sales.

The sales represented by Figure 3 shows low-peaks at the end of each year, which is explained because the stores are closed at New years. Moreover, we can observe a pattern at each year, with increased sales at the end of the year which overlaps with the Christmas eves, suggesting a seasonal pattern. These patterns are highlighted after smoothing the sales data with a 7 days moving average (the continuous red lines from the Figure 3).

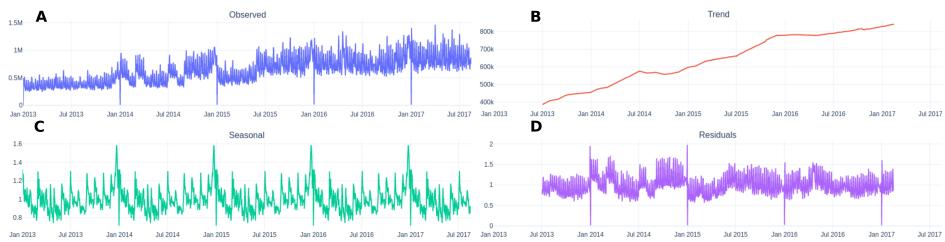


**Figure 3: Target time series data.** The y-axis represents the training sales from 2013-01-01 to 2017-08-15, and the x-axis shows the time in days (1,087 days). The sales were aggregated by all stores and products (see Figure 1). The raw data, and smoothed data (7 days moving average) are denoted by the continuous blue, and red lines, respectively.

### III.1. Time series Decomposition

Time series are defined as a sequence of data organized in time order. The key components of time-series are: trend, seasonality, and residuals, displayed in Figure 4B, Figure 4C, and Figure 4D, respectively. Trend is the long-term direction of the time series, seasonality describes the periodic behavior such as holidays, and residuals represent the irregular fluctuations that we are not able to predict using the trend and/or seasonality.

In our target time series, we observe an upward trend, with a clear seasonality factor (this will be further analyzed below), and the residuals do not follow a random pattern. These insights will be used to select an adequate forecast model.



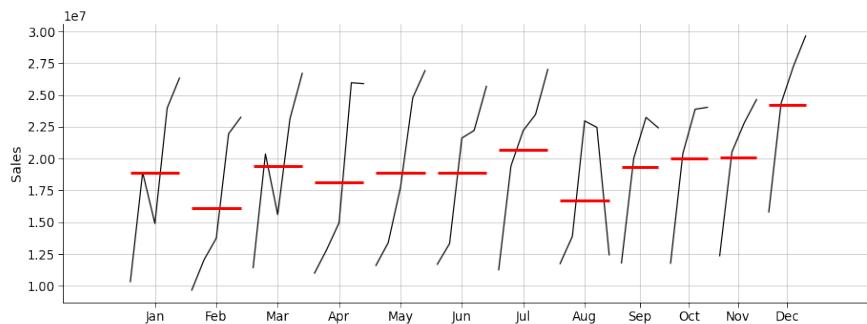
**Figure 4: Time series decomposition.** (A) Raw sales time series (sames as Figure 3). (B) Trend of sales. (C) Seasonality of sales. (D) Residuals of sales.

## III.2. Stationarity and seasonality analyses

The stationary and seasonality are relevant components to select the adequate machine learning model (such as ARIMA, SARIMA, etc.) to generate reliable forecasting. A stationary time series is defined, if their statistical properties such as mean, and variance are all constant, and independent of time.

In consequence, we implemented a *Dickey-Fuller test* to assess stationarity. We obtained a *p-value* of 0.09, using a significance level of 0.05, we can reject the null hypothesis (the series is stationary) concluding that our series is non-stationarity.

In terms of seasonality, we aggregate the sales for each month and we can clearly see on average an increase of sales during December (Figure 5), which makes sense because Christmas and New years represent a season of the year with more transactions in other retail businesses. Furthermore, we highlighted February as the lowest sales month.



**Figure 5: Analysis of seasonality.** The y-axis represents the sales aggregated by month, x-axis shows the analyzed months (from January to December), the horizontal red lines denote the sales average by month (55.5 months).

---

# Appendix

## IV

---

## Supplementary information

This work was written with emacs<sup>1</sup> using L<sup>A</sup>T<sub>E</sub>X<sup>2</sup>, using only **Free and Open Source software**. All the computational analysis were carried out using Linux-based distributions. The figures were generated with Python (matplotlib<sup>3</sup>/seaborn<sup>4</sup>/plotly<sup>5</sup>) and Inkscape<sup>6</sup>.

Contact: [razielar@gmail.com](mailto:razielar@gmail.com)

---

<sup>1</sup><https://www.gnu.org/software/emacs/>

<sup>2</sup><https://www.latex-project.org/>

<sup>3</sup><https://matplotlib.org/>

<sup>4</sup><https://seaborn.pydata.org/>

<sup>5</sup><https://plotly.com/>

<sup>6</sup><https://inkscape.org/>